Online Journal of Communication and Media Technologies, 2023, **13**(4), e202349
e-ISSN: 1986-3497

OPEN ACCESS

Research Article

Characterizing gender stereotypes in popular fiction: A machine learning approach

Chengyue Zhang [1*]
0009-0007-4466-8845

Ben Wu [2]
0009-0005-6260-8999

[1] Phillips Exeter Academy, Exeter, NH, USA
[2] University of California, Riverside, CA, USA
[*] Corresponding author: chengyuelunazhang@gmail.com

Citation: Zhang, C., & Wu, B. (2023). Characterizing gender stereotypes in popular fiction: A machine learning approach. *Online Journal of Communication and Media Technologies, 13*(4), e202349. https://doi.org/10.30935/ojcmt/13644

ARTICLE INFO

Received: 14 May 2023

Accepted: 24 Aug 2023

ABSTRACT

Gender representation portrayed in popular mass media is known to reflect and reinforce societal gender stereotypes. This research uses two methods of natural language processing–Word2Vec and bidirectional encoder representations from transformers (BERT) model–to analyze gender representation in popular fiction and quantify gender bias with gender bias score. Word2Vec, which represents the words in vectorized format, can capture implicit human gender bias with the geometry relationship between word vectors. BERT, a newer pre-trained deep learning model, is specialized in understanding words in the larger context it appears in. The research will compare the results obtained from Word2Vec and BERT. With book check out records from the Seattle Public Library checkout dataset–an ongoing open source dataset from the public library system of Seattle, WA–the research aims to identify evolutionary trends of gender bias in popular fiction and analyze consumer preferences regarding gender representation.

Keywords: gender bias, gender representation, popular fiction, public reading interests, word embedding, BERT

INTRODUCTION

Media serves as a filtered lens through which individuals interpret reality, significantly influencing how consumers perceive themselves and the world around them (Castañeda, 2018). Prolonged exposure to negative or inaccurate portrayals of minority groups in the media can foster stereotypes in consumers' minds (Kidd, 2016). In 2020, Dixon-Fyle et al. (2020) conducted a study tracking data from 365 large US- and UK-based companies since 2014, along with 1,039 additional companies across 15 countries over five years, finding that organizations with more diverse leadership were 12% more likely to outperform their counterparts. Consequently, the representation of minority groups in the media has garnered global attention. For instance, US non-profit organizations like GLAAD (Gay & Lesbian Alliance Against Defamation) release annual reports concerning media representation of minority groups, specifically addressing the intersectionality of minority identities. In the Middle East and North Africa, private museums established by historical minorities strive to diversify historical narratives and challenge the homogeneous accounts presented by most regional public museums (Rey et al., 2020). While diverse media representation is insufficient on its own, it constitutes an initial step toward fostering a more inclusive society, where individuals feel universally accepted (Dixon-Fyle et al., 2020).

Books, as hedonic (non-utilitarian) goods, are highly reflective of public opinions. Hedonic, as defined by Khan et al. (2005), refers to items possessing entertainment value but not deemed essential. Although consumption of popular media is not crucial for survival, it is often justified by a strong preference for less hedonic alternatives (Khan et al., 2005). Since hedonic goods carry a high consumption risk, their consumption motives are closely tied to emotions and may even express the consumer's individuality and symbolic character (Clement et al., 2006). Moreover, the consumption of books is susceptible to external social influences, such as social media and current events. Consequently, the consumption of books serves as a valuable proxy for assessing societal preferences for cultural goods.

Natural language processing (NLP) models are trained on text various text corpora to understand natural language, and such models are shown to exhibit social bias (Babaeianjelodar et al., 2020; Bolukbasi et al., 2016; Caliskan et al., 2017). Our research uses two commonly used NLP techniques, Word2Vec and bidirectional encoder representations from transformers (BERT), to measure gender bias within popular fiction. Word2Vec represents words in vectorized format and is proven to be capable of capturing implicit human gender bias with the geometry relationship between word vectors (Bolukbasi et al., 2016; Caliskan et al., 2017). On the other side, BERT is a newer pre-trained deep learning model specialized in understanding words in the larger context it appears in (Devlin et al., 2018). We compare the results by both methods and determine their mutual validation.

Within the field of bias in NLP, there is a lack of understanding of the social impact of bias in NLP (Blodgett et al., 2020; Delobelle et al., 2021). Although many researchers examine and mitigate bias in such NLP models, the loosely defined intrinsic bias measured in NLP models does not translate to extrinsic bias in the performance of their downstream tasks or the actual social impact of NLP models (Blodgett et al., 2020; Goldfarb-Tarrant et al., 2020). To address the challenge, our research takes a less-researched route of using the bias in NLP model to measure the bias in the real world. We aim to

(1) quantify the evolution of gender bias in popular fiction and

(2) investigate societal gender bias using consumer preference of popular fiction with different degrees of gender bias.

Consumer preference is indicated by the number of times a book is checked out. By incorporating the consumption/demand side of the data, our research emphasized the real-world impact of gender bias in popular fiction. This demand data reflects

(1) how much the book along with the bias it contains spread in society and from the opposite perspective and

(2) the consumer's preference for fiction in terms of gender bias.

Then, aggregating the calculated gender bias of a book by its demand, a more comprehensive understanding of each book's societal impact and societal gender bias is obtained. Our approach provides a framework for analyzing media consumption patterns alongside gender bias measurements to better bridge the gap between bias examined by NLP models and the social impact of such bias.

Our research has several practical implications. Firstly, the quantitative data on gender bias in contemporary popular fiction that our research produces can help consumers make better-informed judgments on the potential gender bias in the books they consume. Furthermore, the method and results of this research can aid content creators, distributors, and policymakers in identifying areas, where gender representation in media could be improved and promote more balanced and inclusive media representation. Clearly illustrating the extent to which gender bias in a certain book spreads, our findings urge writers and distributors to be aware of the societal impact of their product. Policymakers can use our method to monitor the evolving landscape of gender bias in fiction and effectively target specific books and material types that significantly influence societal gender bias (**Figure 1**).

LITERATURE REVIEW

The current literature review contains two sections: biases in cultural goods, traditional method(s) for analyzing bias in cultural goods, and gender biases in NLP. In the first section, we will first discuss the general history of representational bias in media and how the stereotypes created impact minority groups, which serves as the motivation for this paper. The second section discusses relevant research that aims to measure, mitigate, and analyze biases in NLP. These research provide methods applicable to our research on measuring the evolution of biases in books over time.

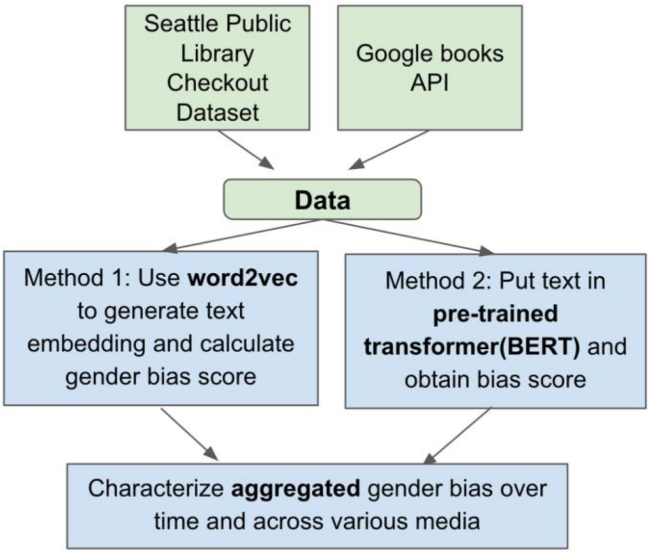
Seattle Public Library Checkout Dataset
Google books API
Data
Method 1: Use word2vec to generate text embedding and calculate gender bias score
Method 2: Put text in pre-trained transformer(BERT) and obtain bias score
Characterize aggregated gender bias over time and across various media

Figure 1. Flowchart of data collection & method of the research (Source: Authors)

Biases in Culture Goods and Its Impact

In the technological world, people are surrounded by all kinds of media and influenced by this large amount of media input. With newspapers being one of the earliest forms of media, the media has a history of enforcing national rules and solidifying national identities by excluding the "others" (Bleich et al., 2015; Waisbord, 2004). Early researchers in media studies believed that media created by humans is reflective of society (Beltrán, 2018; Fürsich, 2010). In the 1960s and 1970s, as television sets became prevalent in every household in the US, scholars, and politicians slowly realized the power of such mass media in influencing its audiences (Dahlgren, 2000; Smith & Granados, 2009). In 1997, Hall expanded the role of media, stating that representation in media not only reflects but also creates shared cultural meanings and connotations within society (Hall, 1997).

Media, by its definition, amplifies certain voices. Due to the imbalance of power within society, people in positions of power enjoy more opportunities to influence media and are more likely to be represented in media (Beukeboom & Burgers, 2019; Leavitt et al., 2015; Shrum, 2009). Thus, negative biases of minority groups held by people in power will be amplified by the media and result in a bigger group of people, consciously or unconsciously, believing in the biases (Brooks & Hébert, 2006; Fiske, 1993). Because of the cyclical nature of the creation of biases, it is hard to pinpoint whether the representation causes biases or if biases result in inaccurate representations (Dev et al., 2021; Fast et al., 2016). Either way, analyzing representation in media can serve as a tool to measure biases in the real world, and improving the representation of minorities in media serves a crucial role in combating discrimination.

In *The Danger of a Single Story*–a TED talk given in 2009 by Nigerian writer Chimamanda Adichie about her experience growing up in Nigeria and studying in college in the US–Adichie explained that inaccurate and one-dimensional representation in media causes ignorance and monolithic prejudices, which over time develop into societal stereotypes (Adichie, 2009). As the stereotypes repeatedly occur in media, they become entrenched in the social consciousness. These stereotypes contribute to bias by perpetuating generalized beliefs about minority groups, justifying discriminatory actions towards minority groups, or even influencing the minority group's self-perception (Shrum, 2009; Zastrow et al., 2019). For example, the prevalent depiction of transgendered people as comedic relief characters not to be taken seriously in American media since the 1970s leads to disproportionate harassment and hate crimes against transgender people on the streets (James et al., 2015; Kearl, 2014; McInroy & Craig, 2016). Similarly, the lack of contemporary representation of Native Americans in American media limits the self-understandings of Native Americans, leading to low self-esteem and lack of belonging (Fryberg et al.. 2008; Leavitt et al., 2015).

Fiction, as a cultural good specifically, has a special impact on the perception and spread of biases (Glick & Fiske, 2011). Before the existence of digital media, humans used printed text to express their ideas. Among all writing genres, fiction communicates with easy-to-understand stories and is thus the most accessible and influential (Johnson et al., 2014; Mar & Oatley, 2008). According to research by Green et al. (2003), readers exhibit "low elaborative scrutiny and high experienced transportation" when consuming fiction. The element of story-telling and narrative in fiction tabs on human's ability to empathize with the characters without the critical thinking usually exhibited when consuming academic and non-fictional works (Amossy & Heidingsfeld, 1984). Even successful political leaders use narrative power to persuade and gain support from their people (Hanne, 1994). Famous fiction like *Brave new world* and *Animal farm* are especially known for their powerful display of political opinion and social criticism through storytelling (Booker, 1994; Clapper, 1995).

Traditional Method for Analyzing Bias in Media

Biased gender representation in fiction has been prevalent throughout history (West, 2010). In previous studies regarding gender representation in fiction, researchers usually manually identify the characterization of characters of different genders (Crabbs & Bielawski, 1994; Hamilton et al., 2006; Hubler, 2000). However, because manual analysis of gender bias is very labor intensive, these research usually only examine limited corpora, such as one particular book or the complete work of a certain author (Ochieng, 2012). Modern researchers have employed modern technology to their benefit to further investigate gender biases within texts. Further explained in the next section, NLP allows researchers to process texts on a large scale (Kraicer & Piper, 2019; Merrick, 2012). Existing research that uses NLP to analyze fiction focuses on character characterization. Bamman et al. (2014a), for example, examined 15,099 English novels from 1700 to 1899 to identify various character types. NLP model created by Bamman et al. (2014a) was then used by Kraicer and Piper (2019) and Underwood et al. (2018) to analyze the depiction of different genders in modern fiction. Fast et al. (2016) also use NLP to study the relationship between gender representations in books and the gender of the authors.

Natural Language Processing and Gender Bias in Natural Language Processing

As established previously, representation in media is a prominent source of discrimination. Still, it is also hard to track that influence due to the massive amount of information consumed by individuals nowadays, such as text, video, music, and images. These sources of information are hard to describe and quantify with straightforward values and numbers (Brooks & Hébert, 2006; Costa-jussà, 2019). In analyzing texts and human language specifically, NLP uses machine learning to quantify natural language to evaluate the meaning of human speech or texts. Biases are found in various parts of an NLP system, such as training data, research design, pre-trained model, and algorithm (Hovy & Prabhumoye, 2021; Sun et al., 2019). The presence of biases within any of these parts will cause biased results from NLP model and thus produce bias in the downstream tasks of NLP, which include machine translation, speech recognition, and sentiment analysis (Du et al., 2021; Zhao et al., 2017). Biases in NLP can be categorized into allocation bias and representation bias: allocation bias means that an NLP model wields better performance on data associated with the majority group, and representation bias means that NLP models capture stereotypes within its association for identity groups (Burns et al., 2019; Sun et al., 2019).

In terms of gender bias within NLP, many researchers aim to measure gender bias through "psychological tests, performance differences between genders for various tasks, and the geometry of vector spaces" (Caliskan et al., 2017; Sun et al., 2019). Although the existence of gender bias within NLP is proven, there are still many problems within the area of study, First, bias analysis of NLP tools is unbalanced and incomparable across languages (Chen et al., 2021; Matthews et al., 2021). Second, there are very few research that acknowledge non-binary gender or calculate the biases toward transgender people (a more detailed discussion regarding this topic is given later). Finally, although gender bias in NLP is widely acknowledged, there is a lack of research to find a solution to the problem (Costa-jussà, 2019; Stanczak & Augenstein, 2021). The few common ways to mitigate gender biases are word embedding Debias and Counterfactual data augmentation (Lu et al., 2020; Maudslay et al., 2020).

Besides studies that investigate the gender bias within NLP models and their downstream tasks, research also use NLP to investigate gender bias with text data (Stanczak & Augenstein, 2021). Many researches use NLP techniques to identify characteristics commonly associated with different identities: Chapman et al. (2020), Hagiwara et al. (2017), and Li et al. (2017) use linguistic inquiry and word count (LIWC) to quantify linguistic and psychological characteristics in text data describing gender and racial groups; Bamman et al. (2014a) use entity-centric modeling to categorize fictional characters into latent persona and calculate the proportion of character's gender in each persona category; Fast et al. (2016), Otterbacher (2015), and Wagner et al. (2015) uses crowd-sourced lexical categories to identify gender stereotypes. Starting with Garg et al. (32018) and Hamilton et al. (2018), however, research adopted a new approach of using measured bias with word embedding as an indication bias in the real world. This new method allows research to incorporate time dimensions when measuring biases within the text. Hamilton et al. (2018) trained word embedding using data from each decade from 1980 to 2012 separately. By comparing these embeddings based on texts from different time periods, they found changes in words' semantic meanings over time. Garg et al. (2018) used three pre-trained data sets to compute the gender bias and ethnic bias in 20th and 21st-century America. The calculated bias corresponded with occupational consensus data during the time, proving that bias in word embeddings mirror social change. Based on Garg et al. (2018) and Hamilton et al. (2018), more recent studies use NLP to analyze bias in songs (Boghrati & Berger, 2023) and movies (Khadilkar et al., 2022) over time by examining gender-related vector association in word embedding.

DATA

Overview

This research uses the checkout history of Seattle Public Library (https://data.seattle.gov/Community/Checkouts-by-Title/tmmm-ytt6) from 2004 to 2021. We took the top 0.1% of most checked-out fiction and, extract from Google Books' API's metadata its summary. The number of times a title is checked out during each month is used for aggregation to measure the total societal gender bias contributed by the book. The checkout number of a book is timed with the book's measured bias to obtain the book's aggregated bias. The aggregated bias considers that bias within a piece of popular media will be spread differently due to the amount of time that the book is consumed.

Seattle Public Library checkout dataset is an ongoing public dataset containing every checkout entry from Seattle Public Library, comprising 27 branches in various neighborhoods around the city, from April 2005 to the present. Compared to other possible measurements for the consumption of books, such as selling records from book vendors or checkout records from private libraries or school libraries, public library checkout records can more accurately reflect the consumption preferences of society as it is relatively more accessible to the general public. Although the reading preferences embodied by the dataset do not wholly reflect the US consumers' preferences for books, it is the only publicly available checkout record from all US public libraries.

The dataset is updated monthly, and when we downloaded the data from the official website in March of 2022, the dataset had a total of 39,692,744 entries. Each entry includes 12 columns: UsageClass, CheckoutType, MaterialType, CheckoutYear, CheckoutMonth, Checkouts, Title, ISBN, Creator, Subjects, Publisher, and Publication Year. UsageClass indicates whether the item is 'physical' or 'digital.' CheckoutType indicates the vendor tool (Horizon, OverDrive, Freegal, Hoopla, and Zinio) used to check out the item. MaterialType indicates whether the item is a book, song, movie, music, magazine et cetera. ISBN, which stands for international standard book number, is a numeric commercial book identifier, and in the dataset, over 99% of entries are missing this data. Therefore, we use the book title and author combination to identify each book instead. **Table 1** shows a few sample entries.

Data Cleaning

A series of data-cleaning jobs are performed on the book dataset. First, laptop checkouts, uncataloged folder or bag checkouts, and entries with missing or unknown titles are eliminated. Then, MaterialTypes were constrained to only 'BOOK,' 'EBOOK,' and 'AUDIOBOOK' to exclude all other entries of music discs and movies. Next, we selected entries with subjects that contain the keyword 'fiction.' After constraining the Subject and MaterialType columns, there are 12,331,217 entries left in the dataset, including 6,637,661 physical book entries, 3,996,189 ebook entries, and 1,697,367 audiobook entries.

Next, we decided only to include fictional materials because fiction can effectively reflect social biases. We did so by selecting entries with subjects that contain the keyword 'fiction.' After constraining the Subject and

Table 1. Sample entries from the Seattle Public Library checkout dataset

Usage class	Checkout type	Material type	Checkout year	Checkout month	Checkouts	Title	ISBN	Creator	Subjects	Publisher
Digital	Over drive	Audio book	2020	6	4,903	So you want to talk about race (unabridged)		Ijeoma Oluo	African American nonfiction, nonfiction, politics, & sociology	Blackstone Audio, Inc.
Physical	Horizon	Mixed	2018	10	2,073	FlexTech--Laptops			Laptop computers, iPad computer, & tablet computers	Lenovo,
Digital	Over drive	E-book	2021	7	1,088	The vanishing half: A novel		Brit Bennett	Fiction & literature	Penguin Group (USA), Inc.
Physical	Horizon	Book	2019	1	801	Becoming/ Michelle Obama		Obama, Michelle, 1964-	Obama Michelle 1964, presidents spouses US biography, African American women lawyers Illinois Chicago biography, legislators spouses US biography, & autobiographies	Crown
Digital	Over drive	Audio book	2020	4	1,894	Harry Potter & the sorcerer's stone: Harry Potter series, book 1 (unabridged) (unabridged)		J. K. Rowling	Juvenile fiction & juvenile literature	Pottermore
Digital	Zinio	Magazine	2019	2	996	The New Yorker				

MaterialType columns, there are 12,331,217 entries left in the dataset, including 6,637,661 physical book entries, 3,996,189 ebook entries, and 1,697,367 audiobook entries. Besides the fact that it is technically easier to identify the biases within fiction that do not include many subject-specific vocabularies, there are two more reasons for this decision from the perspectives of the author and the reader.

Firstly, fictions, songs, and movies contain subjective descriptions, which inevitably reflect the biases of the creator and society (Huang et al., 2019). The following quote from The Picture of Dorian Gray, for example, is used to depict the strangeness of Dorian Gray's action after being influenced by Lord Henry's talk of Hedonism: "At another time, he[Dorian Gray] devoted himself entirely to music … slim turbaned Indians blew through long pipes of reed or brass and charmed–or feigned to charm–great hooded snakes and horrible horned adders. The harsh intervals and shrill discords of barbaric music stirred him at times when Schubert's grace, Chopin's beautiful sorrows, and the mighty harmonies of Beethoven himself fell unheeded on his ear." The negative depiction of Indian music and the bias towards Indian culture is not a major message of the book. The word 'Indian' is never mentioned again outside the paragraph in which the quote is located. In other words, Oscar Wilde did not intend to express his negative feelings toward African and Indian music. He simply used a commonly held belief of British society during his time to convey the bizarreness of Dorian Gray's action (Muhammed, 2020). However, this negative and biased description of Indian music enforced and will continue to enforce the societal bias of Indian culture being barbaric and oriental (Amossy & Heidingsfeld,

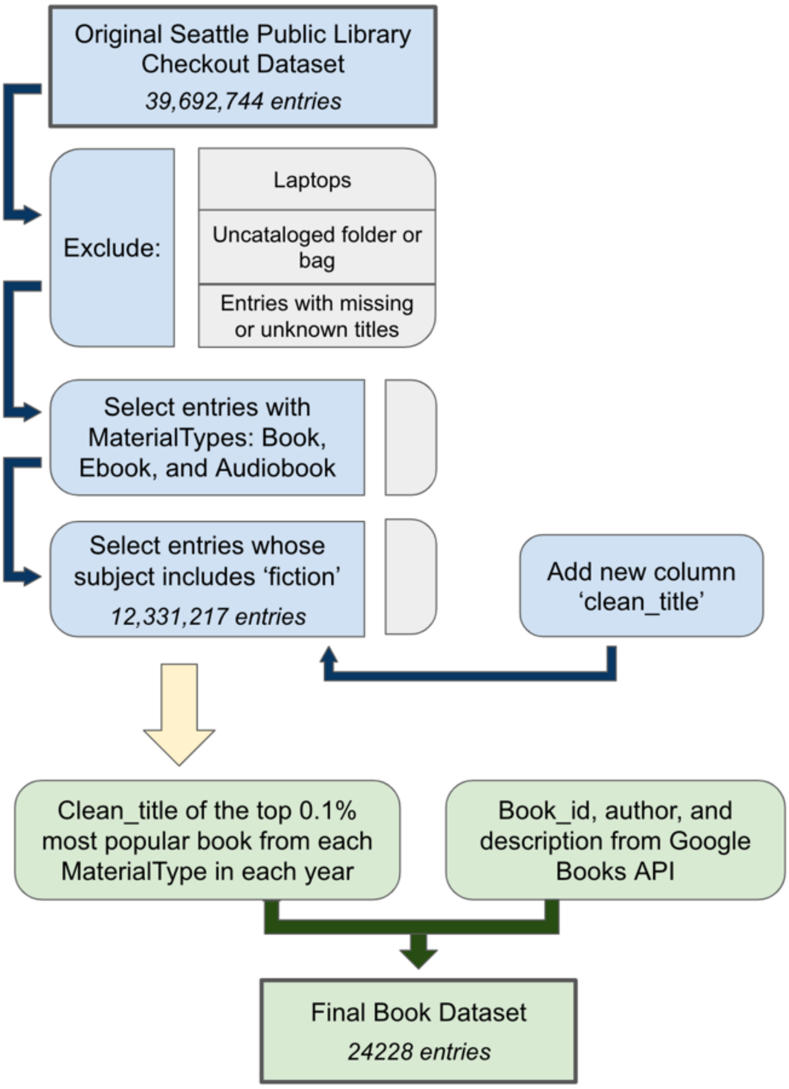
Original Seattle Public Library Checkout Dataset
39,692,744 entries
Exclude:
Laptops
Uncataloged folder or bag
Entries with missing or unknown titles
Select entries with MaterialTypes: Book, Ebook, and Audiobook
Select entries whose subject includes 'fiction'
12,331,217 entries
Add new column 'clean_title'
Clean_title of the top 0.1% most popular book from each MaterialType in each year
Book_id, author, and description from Google Books API
Final Book Dataset
24228 entries

Figure 2. Flowchart of the data cleaning process for books (Source: Authors)

1984; Said, 2016). In other words, fiction reflects the norms generally accepted in the environment in which the author lives and the author's personal and biased opinions. Similar examples can be found in songs and movies as well: in classic Hollywood movies of the 1910s and 1920s, women were often portrayed as either mothers or sexual objects, coinciding with the general unequal treatment of women at the time (Smelik, 2007).

Secondly, popular fictions match the readers' consumption preferences. People read fiction to be entertained and not challenged. So when choosing which book to read, readers tend to select the ones that fit their preferences and are easy to relate to. Only if a reader can understand and empathize with the character will they keep on reading (Green et al., 2003), and this means that biases in the fiction largely match the biases of the reader. Besides, readers are easily influenced when reading fiction because they empathize with the character's experience instead of analyzing the plot and each sentence critically as they would perhaps a non-fictional work. Thus, the biases in the book that the readers did not originally possess are likely to be taken on by the readers (**Figure 2**).

As for some popular or classic books published multiple times by different publishers in different material types, there are slight variations in titles. The same book is recorded under different titles with small variations (an example of which is shown in **Table 2**). To combine them, we changed the edited title to all lowercase and deleted text like '(unabridged)' and '/' plus the author's name. The newly edited titles are stored in a new column called 'clean_title,' leaving the original title intact.

Table 2. A sample of different editions of *the Picture of Dorian Gray* in the dataset

No	Title	Material type	Publisher	clean_title	Checkouts
1	The Picture of Dorian Gray/Oscar Wilde	Book	Penguin,	The Picture of Dorian Gray	997
2	The Picture of Dorian Gray	E-book	Duke Classics	The Picture of Dorian Gray	712
3	The Picture of Dorian Gray (unabridged)	Audio book	Blackstone Audio, Inc.	The Picture of Dorian Gray	529
4	The Picture of Dorian Gray (unabridged)	Audio book	Tantor Media, Inc	The Picture of Dorian Gray	383
5	The Picture of Dorian Gray	E-book	Random House, Inc.	The Picture of Dorian Gray	351
6	The Picture of Dorian Gray: An annotated, uncensored edi …	Book	Belknap Press of Harvard University Press	The Picture of Dorian Gray	219

Table 3. Samples of book descriptions

Book_id	clean_title	Author	Description
C4BhbzcOIAMC	tangerine	['Edward Bloor']	12-year-old Paul, who lives in shadow of his football hero brother Erik, fights for right to play soccer despite his near blindness & slowly begins to remember incident that damaged his eyesight. An ALA best book for young adults. Reprint. Jr Lib Guild.
yDGgDwAAQBAJ	the witch elm	['Tana French']	Named a New York Times notable book of 2018 & a best book of 2018 by NPR, The New York Times book review, Amazon, The Boston Globe, LitHub, Vulture, Slate, Elle, Vox, and Electric Literature "Tana French's best and most intricately nuanced novel yet." —The New York Times An "extraordinary" (Stephen King) and "mesmerizing" (LA Times) new standalone novel from the master of crime and suspense and author of the forthcoming novel The Searcher. From the writer who inspires cultic devotion in readers …
0_SQxAEACAAJ	the bride test	['Helen Hoang']	Khai Diep has no feelings. Well he can't feel big emotions like love. He thinks he's defective. His mum knows that his autism means he just processes emotions differently and goes to Vietnam to find him a wife. As a mixed race girl living in Ho Chi Minh City, Esme Tran has always felt out of place, so when…

Next, we took the top 0.1% of the most checked-out books from each MaterialTypes each year and compiled their clean_title, CheckoutYear, MaterialType, and Checkouts into a new file. Then, we used the clean-title column of the selected data to extract metadata regarding the book, including its book id, description, and author, from Google Books API. The data extracted from Google Books (book id, description, and author) and the columns clean_title from checkout data are combined into one file. There are 2019 rows of data and a total of 842 different books (an example of which is shown in **Table 3**).

Data Analysis

Figure 3 is graphed using the cleaned data, and the entries from 2022 are excluded because we only had data from the first three months of 2022. The number of total checkout entries increased from 2005 to 2019 and drastically dropped in 2020 when COVID-19 hit the US, interrupting the borrowing of physical books. Looking at the three material types specifically, the number of times that audiobooks and ebooks were borrowed increased steadily from 2005 to 2021, and only the borrowing of physical books (labeled book in data) is significantly affected by the surge of COVID-19 in 2020. We can also notice that the number of physical books borrowed has been slowly decreasing since 2010. Therefore, the overall increase in the number of books borrowed is only due to the increased number of audiobooks and ebooks borrowed.

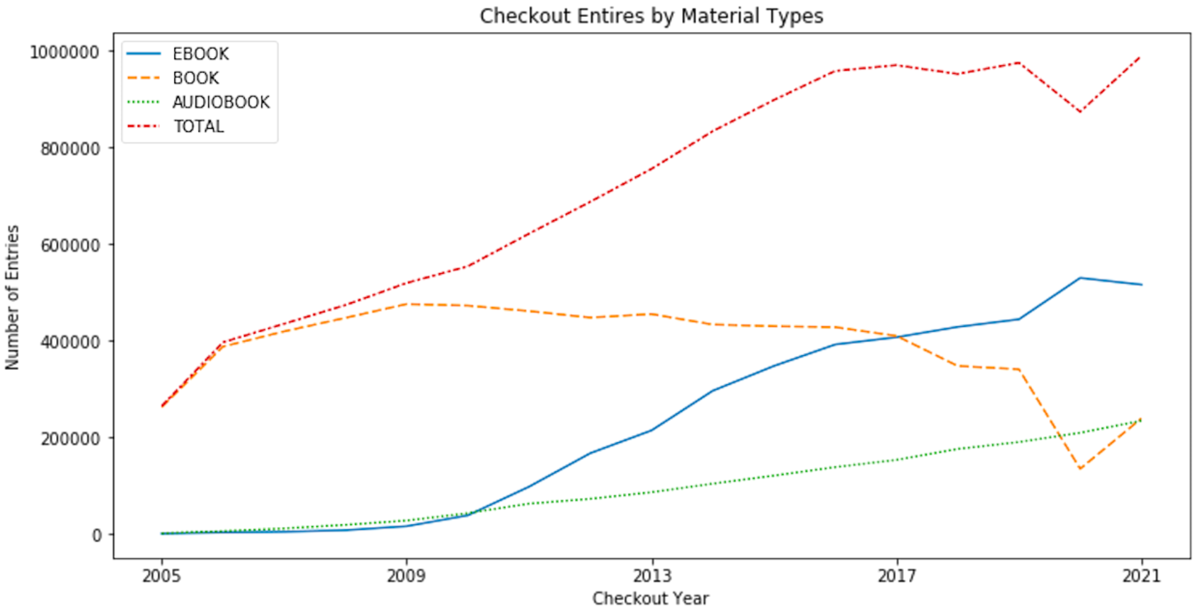

Figure 3. Checkout entries by material type from 2005 to 2021 (Source: Authors)

Table 4. Top-15 most popular books yearly from 2015 to 2021

No	clean_title	CheckoutYear	Checkouts
1	The vanishing half	2021	16,720
2	Where the crawdads sing	2019	15,447
3	Where the crawdads sing	2020	11,594
4	The midnight library	2021	10,070
5	All the light we cannot see	2015	9,146
6	Anxious people	2021	8,478
7	Crazy rich Asians	2019	8,225
8	Little fires everywhere	2018	8,190
9	The girl on the train	2015	7,863
10	Crazy rich Asians	2018	7,382
11	The underground railroad	2017	7,352
12	The power	2018	7,249
13	The Dutch house	2020	7,113
14	All the light we cannot see	2016	7,024
15	Little fires everywhere	2019	6,872

Table 4 shows the top-15 most popular (borrowed the most times considering all material types) books each year from 2015 to 2021, as well as the number of times they were checked out that year. All 15 books were checked out after or during 2015 and aligned with best-sellers lists from the press.

METHOD

Word Embedding Model

Basic model

Word Embedding is a framework for presenting words in vector forms as a byproduct of the process of NLP (Mikolov et al., 2013a).

Before the invention of word embedding, people used statistical measures to analyze how important a word is in a given text (Zhang et al., 2011). Word Embedding models use a three-layer Neural Network, including an input layer, hidden layer, and output layer, to learn the relationship between words from given texts through various method models, producing the word embedding for every word in the given texts in the process. The number of nodes in the hidden layer matches the number of dimensions the word embedding possesses. After NLP model is trained and performs well enough for the given task, the word embedding of each word can be taken from the hidden layer of the neutral network. multiple neural network models were developed to generate word embedding. The two most common models are continuous-bag-of-words (CBoW) and skip-gram. In CBoW, the neural network learns to predict the masked word in a sentence from its surrounding words (Mikolov et al., 2013a). Skip-gram, on the other hand, performs the task of predicting multiple words next to one single given word (Guthrie et al., 2006; Mikolov et al., 2013b). Different from CBowW and skip-gram, which studies the relationship between the target word and its local context, GloVe (global vectors for word representation) looks at a larger window and studies the global corpus statistics of words co-occurrence by recording the number of times words appear within a certain-size window surrounding the target word (Pennington et al., 2014). The three models mentioned above produce static word embeddings, which can be used for downstream tasks, such as language modeling, sentiment analysis, and machine translation (Dev et al., 2021; Wang et al., 2019). They perform the task of Word2Vec, representing each word with a multidimensional vector that indicates the complex relationship of words within the word list.

Word embeddings are proven to accurately reflect biases in society. Bolukbasi et al. (2016) demonstrated gender bias in word embedding with a geometric approach to the word embedding system. They found a sexist association established through word embedding: 'man' to 'computer programmer' as 'woman' to 'homemaker.' Caliskan et al. (2017) compared the biases found in static Google News embedding with various biases shown in society through implicit association test (IAT). Caliskan et al. (2017) found that biases in word embedding reflect not only problematic biases, such as gender biases and racial biases but also neutral biases, such as the association of flowers with pleasantness and insect with unpleasantness.

Though biases in word embedding pose problems when used downstream (Basta et al., 2019; Pennington et al., 2014), this research, however, uses word embedding to measure biases in books and thus within society. In this research, we use pre-trained word embedding created by Hamilton et al. (2018) and used by Garg et al. (2018). Hamilton et al. (2018) trained word embedding from six datasets from Google Books with data from 1800 to 1999 to measure the semantic change in over 30,000 words across four languages. Each dataset is separated by the publication year of the text, and the texts from the same decade are grouped and used to train word embedding. The word embedding we used is trained by the skip-gram model on English fiction written in the 1990s recorded in Google Books. It consists of 100,000 words and 300 dimensions.

The primary reason for using a pre-trained model is that our research does not have a large enough text corpus to train an effective word embedding model ourselves. Constructing a high-quality word embedding model requires an extensive collection of text data (Lai et al., 2016; Rezaeinia et al., 2019), which is not readily accessible or feasible to compile for our specific research focus on published popular fiction due to copyright issues. As a result, we turn to pre-trained models.

In selecting the pre-trained word embedding, it is crucial to consider the compatibility between the corpus used for training the word embedding and the specific task for which it will be employed, as this significantly impacts the embedding's performance (Dhingra et al., 2017; Lai et al., 2016). To address this concern, we carefully selected a word embedding model trained by Hamilton et al. (2018) that aligns best with our research objectives. Many available word embedding models are trained on text data from academic settings, such as Wikipedia and Google News (Dhingra et al., 2017). In contrast, Hamilton et al.'s embedding model is specifically trained using English fiction books from the 1990s, which is compatible with our goal of measuring gender bias in fictional works. By using an embedding model trained on a corpus consisting of fictional literature, we can capture the specific linguistic nuances and gender portrayals prevalent in that domain. However, Hamilton et al.'s pre-trained model has temporal limitations. Because the embedding is trained only on 1990s data, it does not account for new vocabulary that emerge since then or semantic changes of existing vocabulary, which can affect the accuracy of our result.

Calculating book embedding

Word embedding represents each word with a high-dimension vector. To analyze the text, text input must be represented as a vector. Proposed by Salton and Buckley (1988), bag of words (BOW) produces document embedding by simply calculating the weighted average embedding of all or some of the words that occurred in the document. Although simple and proven effective, BOW fails to consider word order (Le & Mikolov, 2014; Wu et al., 2018). Doc2Vec is a more elaborate method that uses an unsupervised framework to learn feature representations from texts (Le & Mikolov, 2014). Still, the effectiveness of Doc2Vec is not conclusive across research (Lau & Baldwin, 2016). There are still other methods, such as paragraph vectors (Dai et al., 2015) and word mover's distance (Kusner et al., 2015), that produce document embedding, and there is no absolute answer regarding which one is the best (Lau & Baldwin, 2016). The embedding and gender bias of songs and movies can be calculated using the described method for books written below.

Following the method in Garg et al.'s (2018) paper, we calculated book embedding based on the average embedding of all the words in the summary of the book. Although not as vigorous as Doc2Vec, it is a simple but effective method to find the embedding of a series of words. The equation we used is shown below. In the equation, $\vec{V}_{booki}$ it represents the embedding of the *i^th^* book in the booklist; $\vec{V}_{win}$ represents the embedding of the *n^th^* valid word in the book description of the *i^th^* book; n_i represents the number of valid words in the description of the *i^th^* book. Individual word embedding of each word found in the word embedding is added. Then, because book descriptions have varying lengths, the sum of word embeddings is divided by the number of valid words used in the calculation to standardize the embedding of each book.

$$\vec{v}_{book_i} = \frac{\vec{v}_{w_{i1}} + \vec{v}_{w_{i2}} + \cdots + \vec{v}_{w_{in}}}{n_i}.$$

Calculating gender bias

Following research by Bolukbasi et al. (2016) and Garg et al. (2018), the bias of a book is defined by the geometric correlation between embeddings: the Euclidean distance between the book embedding and the average embedding of a list of female-indicating words minus the Euclidean distance between the book embedding and the average embedding of a list of male-indicating words. Though seemingly simple, this method measures bias and is very straightforward and grounded in the geometry of embedding vectors (Friedman et al., 2019; Stanczak & Augenstein, 2021).

$$B_{book_i} = \left\| \vec{v}_{book_i} - \vec{v}_{A_f} \right\|_2 - \left\| \vec{v}_{book_i} - \vec{v}_{A_m} \right\|_2.$$

B_{booki} represents the bias of the *i^th^* book; $\vec{V}_{booki}$ represents the embedding of the *i^th^* book; $\vec{V}_{Af}$ and $\vec{V}_{Am}$separately represents the average embedding of a list of female-indicating and male-indicating words(including 'she' and 'he', 'her' and 'his', 'woman' and 'man' et cetera). The Euclidean distance between the book embedding and $\vec{V}_{Af}$ represents how female-leaning the content of the book is. When the distance between the book embedding and $\vec{V}_{Af}$ is small, the book embedding is very similar to $\vec{V}_{Af}$. This likely means that the book has a lot of female representations (Garg et al., 2018). If the book has a female protagonist, for example, the pronouns 'she,' 'her,' and 'hers' will appear many times in the summary, so the book embedding will be similar to the average embedding of the female-indicating word. *The vanishing half* by Brit Bennett, for example, has a cast of mostly female characters, the two Vigne sisters and their daughters, and its bias is calculated to be -0.01474. On the contrary, *American Gods* by Neil Gaiman has mostly a cast of male characters, including Shadow and the mysterious Mr. Wednesday, and its bias is calculated to be 0.02177. The bias of the *i^th^* book is found by taking the difference between the two Euclidean distances, and the negativity of bias indicates how biased the book is. If the bias of a book is more negative, the book is female-leaning and biased toward women. If the bias of a book is more positive, the book is male-leaning and biased towards men.

Calculating Gender Bias With Pre-Trained Transformer

In our second method, we employ the pre-trained BERT model fine-tuned to measure gender bias. BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model developed by Google in 2018 for NLP tasks (Devlin et al., 2018). Unlike the previously used Word2Vec method, which considers each word in isolation, BERT, as a transformer-based model, can understand the context of a text. During the training process, BERT learns to predict missing or masked words in sentences based on the context provided by the surrounding words (Lai et al., 2020). In this study, we utilize the pre-trained BERT base model, which consists of 12 layers and a total of 110 million parameters and has been trained on document-level corpora from English Wikipedia and books (Devlin et al., 2018).

After loading the pre-trained BERT base model, we fine-tune it using the inter-sentence StereoSet dataset (Nadeem et al., 2020, available at https://huggingface.co/datasets/stereoset). Fine-tuning involves training the model on a specific task while leveraging its pre-trained language understanding capabilities. The goal is to adapt the model to the specific nuances and requirements of our research task, which is to identify gender bias in the text (Sun et al., 2019). During the fine-tuning process, the model undergoes supervised learning. It compares its predictions regarding the presence of bias in inputted sentences with the ground truth labels provided in the StereoSet dataset. Through an optimization process, BERT model's parameters are adjusted to minimize the disparity between its predictions and the true labels (Devlin et al., 2018; Merchant et al., 2020; Zhou & Srikumar, 2021). By fine-tuning BERT with StereoSet, we improve the model's performance in recognizing and quantifying gender bias more effectively, leading to a more accurate analysis of gender representation.

The rationale behind selecting the inter-sentence StereoSet dataset for fine-tuning is its unique design and labeling process. StereoSet comprises 17,000 sentences that have been manually labeled by human annotators to indicate the presence of both biased and unbiased content. The dataset is specifically designed to measure and evaluate social biases in pre-trained language models like BERT (Nadeem et al., 2020). By fine-tuning BERT using StereoSet, we aim to enhance the model's ability to identify social biases in text. The inter-sentence nature of StereoSet is also particularly advantageous because it provides a diverse range of inter-sentence data that enable the model to better grasp the contextual information surrounding each word (Dolci, 2022; Nadeem et al., 2020). This context awareness is crucial for accurately detecting and understanding gender bias in text.

Using BERT in conjunction with StereoSet allows for a more accurate and comprehensive analysis of gender bias in popular fiction. This approach helps us better understand the evolution of gender representation in these media types and provides valuable insights into societal preferences and attitudes toward gender representation.

Aggregated Bias

Furthermore, we use the number of times a book is borrowed from the Seattle Public Library to calculate the aggregated bias. Bias within a piece of popular media will be spread differently due to the amount of time that the book is consumed, so the aggregated bias better reflects societal bias by considering consumption.

We timed the number of book checkout entries with the book's bias and got the book's aggregated bias. We calculated the aggregated bias because the bias within a book will be spread differently due to the amount of time that the book is consumed. The aggregated bias will likely be larger if the book is borrowed more times. The aggregated bias better reflects societal bias. The equation B_{aggL} represents the aggregated bias of all books from booklist L; B_{booki} the bias of the *i*th book in booklist L; L represents a book list that contains *i*th books; c_i represents the number of times the *i*th book is checked out.

$$B_{aggL} = \sum_{book_i \in L} B_{book_i} \times c_i.$$

RESULTS

In **Figure 4**, we can see the standard deviation is larger in earlier years and decreases as time progresses. Gender biases in books are becoming less extreme.

In **Figure 5**, similarly, the aggregated bias is leaning more toward women over the years. Comparing **Figure 4** and **Figure 5**, the unaggregated and aggregated biases generally align with each other. The unaggregated bias (**Figure 4**) seems to be "lagging behind" the aggregated bias (**Figure 5**). This trend can be explained by the borrowing pattern of popular books. For example, as seen in **Figure 6**, the book *Gone girls*' (published in June 2012) checkouts peaked in 2013 but continued to be in the top 1% of most checked-out books until 2017. In the unaggregated method, *Gone girl*'s bias will be counted one time in each year from 2012 to 2017, while in the aggregated method, *Gone girl*'s will contribute to the aggregated graph more in prior years and less in the later years, in proportion to its popularity. Thus, the aggregated bias appears to be "lagging behind" the unaggregated bias.

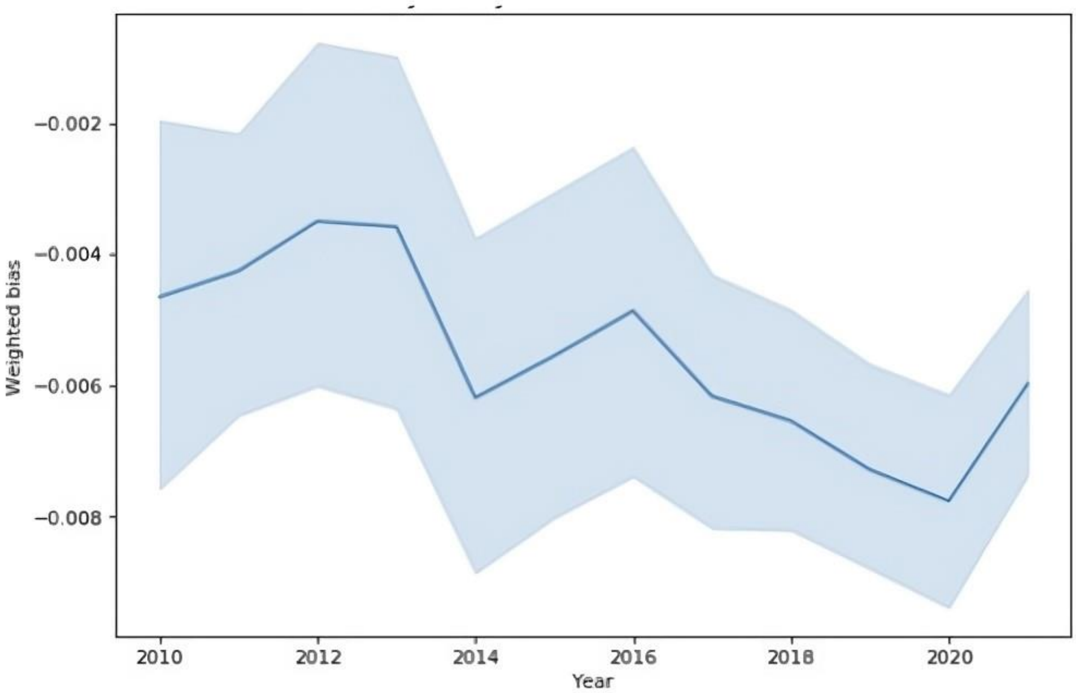

Figure 4. Unaggregated gender bias in books using word2vec from 2010 to 2021 (Source: Authors)

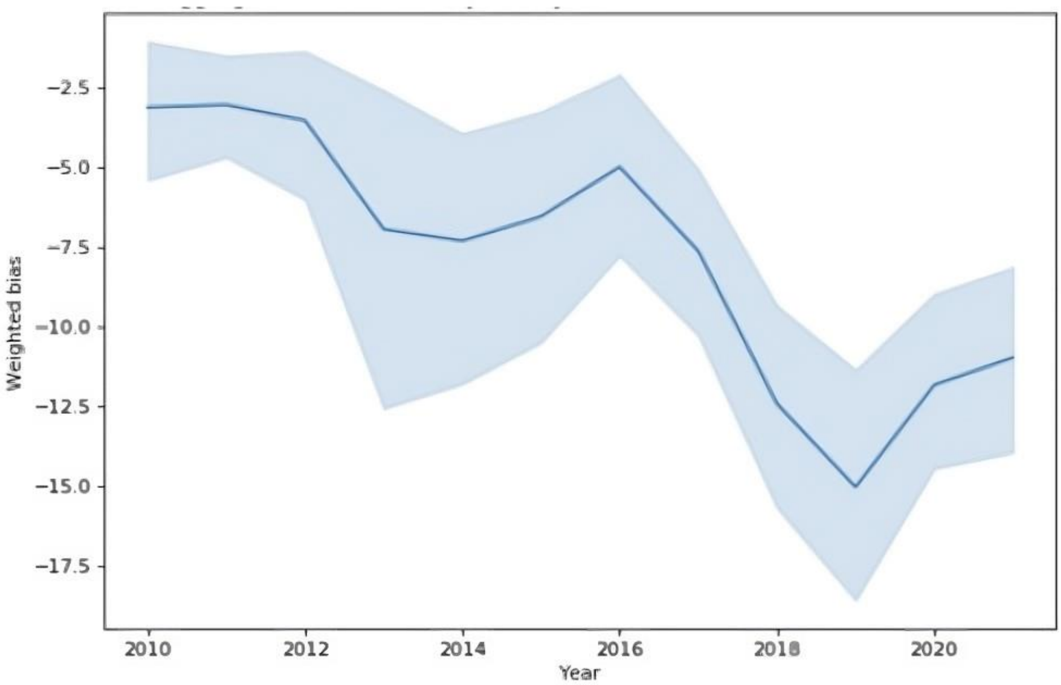

Figure 5. Aggregated gender bias in books using word2vec from 2010 to 2021 (Source: Authors)

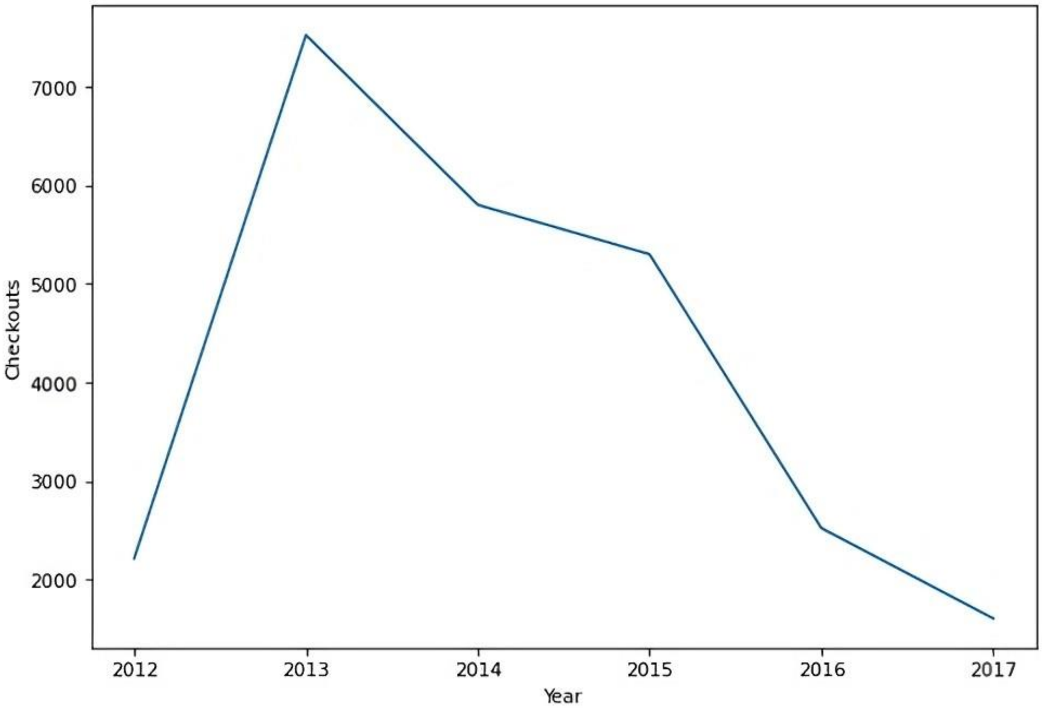

Figure 6. Checkout record of *Gone girl* from 2012 to 2017 (Source: Authors)

Table 5. Top-10 most borrowed books by MaterialType in 2013

No	clean_title	Checkouts	Material Type	Bias_agg	Bias
1	Where'd you go, bernadette	3,977	Physical book	-140.496103	-0.035327
2	Gone girl	3,542	Physical book	-190.608148	-0.053814
3	Gone girl	3,123	E-book	-168.060205	-0.053814
4	Fancy Nancy	2,988	Physical book	-12.595715	-0.004215
5	Fifty shades of grey	2,738	E-book	-57.748195	-0.021091
6	Life after life	2,295	Physical book	-49.697296	-0.021655
7	The casual vacancy	1,994	Physical book	8.893339	0.004460
8	Flight behavior	1,886	Physical book	-38.392757	-0.020357
9	Beautiful ruins	1,843	Physical book	4.826808	0.002619
10	Inferno	1,689	Physical book	28.156936	0.016671

Table 6. Top-10 most borrowed books by MaterialType in 2019

No	clean_title	Checkouts	Material Type	Bias_agg	Bias
1	Where the crawdads sing	6,913	Physical book	-80.030923	-0.011577
2	Normal people	4,431	Physical book	-35.979150	-0.008120
3	The overstory	4,150	Physical book	-11.762703	-0.002834
4	City of girls	3,911	Physical book	-103.413415	-0.026442
5	On earth we're briefly gorgeous	3,478	Physical book	-15.643573	-0.004498
6	Where the crawdads sing	3,294	E-book	-38.134219	-0.011577
7	Big sky	3,214	Physical book	-41.931001	-0.013046
8	Everything I never told you	3,059	E-book	-55.998639	-0.018306
9	Lady in the lake	3,008	Physical book	-75.185779	-0.024995
10	There there	2,977	Physical book	-7.044748	-0.002366

We can also see that the standard deviation of the aggregated bias is significantly smaller than the standard deviation of the unaggregated bias, which aligns with our hypothesis that aggregate bias will lessen the effect of outliers in unaggregated bias. However, around the two local minimum points on the aggregated in 2013 and 2019, the standard deviation of the data increased, indicating that the drop in aggregated bias might be caused by the outliers. In the top-10 most borrowed books in 2013 (**Table 5**), the book *Gone girl* acted as the outlier with a bias score of -0.053814, significantly contributing to the negative gender bias in 2013 in **Figure 7**. In 2019, the drop in aggregated bias is contributed by a few titles, including *City of girls*, *Where the crawdads sing, and Lady in the lake* (**Table 6**).

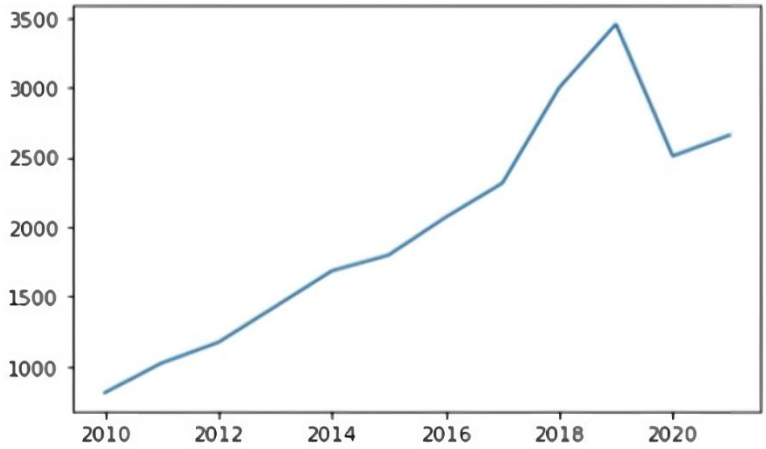

Figure 7. Aggregated gender bias in books using BERT from 2010 to 2021 (Source: Authors)

Additionally, large drop in aggregated bias in the book from 2017 to 2018 might have resulted from the MeToo movement in 2017. This drop occurred in **Figure 4** as well, but it was less significant than in **Figure 6**.

In **Figure 7**, the aggregated bias changes towards having more possibility of being unbiased over the years. Similar to the result from word2Vec, a peak appears during 2019. Though Word2Vec and BERT use distinct mechanics, the results we obtained from the two methods show a similar trend with each other. Their graphs appear as mirrored reflections of each other when placed side by side because smaller values in the word2vec model represent female-leaning content, which corresponds to larger values in BERT model that represent the possibility of being not gender biased.

DISCUSSION

In our research, the two separate NLP methods used to calculate bias produced similar results, which cross validates the general accountability of each other. However, as mentioned in the literary review section, several problems were raised about the study of biases in NLP. Blodgett et al. (2020) summarized the core of these problems nicely: "Even though analyzing 'bias' is an inherently normative process–in which some system behaviors are deemed good and others harmful–papers on "bias" in NLP systems are rife with unstated assumptions about what kinds of system behaviors are harmful, in what ways, to whom, and why." (Blodgett et al., 2020). In response to this problem, the following paragraphs will use specific examples to critically examine how bias is defined in our research, including its advantages and limitations, from three different perspectives.

Concept of Bias and Bias Evaluation

Bias is distinct from accuracy and can be understood as an average deviation from a true value. However, in the context of representation in text, particularly in novels, bias becomes highly subjective and challenging to measure and quantify (Delgado-Rodríguez & Llorca, 2004; Olson, 2012). Bias exists in all forms of human expression, and the concept itself is not necessarily negative, as commonly perceived. For instance, a description of men displaying their emotions through violence may be categorized as biased toward men using our method, potentially reinforcing toxic masculinity and harming men's mental well-being (Kupers, 2005). At the same time, this bias may also affect women who are often portrayed as the recipients of violence, leading to different impacts based on individual experiences (Sculos, 2017). In essence, it is not the bias itself but its impact on human beings that holds significance. However, the same biased content might affect different individuals in distinct ways, further complicating the evaluation of gender bias.

To illustrate the complexities of gender bias in fiction, we examine several examples and discuss the limitations of our method for measuring gender bias. For instance, the novel *Crazy rich Asians* by Kevin Kwan has garnered international popularity and critical attention. While some critics appreciate the portrayal of complex Asian female characters, others argue that the representation in the book perpetuates stereotypes and hegemonic representations. The character Rachel Chu, the protagonist, is depicted as an independent and intelligent woman navigating her relationship and the world of Singaporean elites. While Chu's character exhibits positive attributes, some critics argue that her representation adheres to certain stereotypes, such as the "model minority" trope or the portrayal of Asian women as exotic and submissive (Vijay, 2019; Zhao, 2019).

Our bias calculating system categorizes the book as having a female-leaning bias (with a Word2Vec score of -0.01171 and a BERT score of 1.183846), potentially due to the prominence of female-related terms within the text. However, it is important to note that our model's calculated bias represents a specific perspective based on algorithms and does not provide a definitive answer regarding representation. Critical analysis and understanding of the book's portrayal of characters are necessary to contextualize the calculated bias.

Fifty shades of grey is another controversial case. Some feminist interpretations argue that the book offers women an opportunity to explore their desires and sexual agency, challenging traditional gender roles and expectations (Van Reenen, 2014). However, others criticize the book for reinforcing harmful stereotypes and perpetuating unequal power dynamics between the characters (Bonomi et al., 2013). Our Word2Vec model yielded a slight negative bias of -0.021091, possibly due to recognizing an equal number of descriptions related to both female and male protagonists in the story. BERT model, on the other hand, considers the surrounding text and context, resulting in a score of 2.475644 and a high probability of not having gender biases. NLP models' interpretation offers insights into linguistic patterns present in the summary, but it does not capture the entirety of the book's content or the complexities of its themes. Evaluating gender bias in such cases requires contextual analysis and informed discussions about the book's portrayal of relationships, consent, power dynamics, and feminist perspectives.

The classic novel, *The perks of being a wallflower*, offers valuable insights into our measurement of biases, particularly regarding the traditionally feminine traits exhibited by the main character, Charlie, and their contribution to the calculated bias. Despite being a male protagonist, Charlie demonstrates characteristics typically associated with femininity, such as heightened emotional sensitivity and is described in the summary as an "observant 'wallflower.'" When analyzing this book using the Word2Vec model, we obtained a bias score of -0.00497, while BERT model yielded a score of 0.582947. These scores indicate that the models might have picked up on words traditionally associated with female traits and gave out a slight female-leaning bias score. Charlie's portrayal challenges the conventional expectations and stereotypes associated with male characters. By breaking down rigid gender norms, the book and its portrayal of Charlie help foster a more equal and accepting society that recognizes and celebrates the diversity of human experiences and identities. This subversion of traditional gender roles contributes to the calculated bias in our analysis.

However, our method encounters limitations in the analysis of Margaret Atwood's "The handmaid's tale," a seminal feminist narrative portraying women's resilience in a dystopian society filled with misogyny (Atwood, 2017; Beauchamp, 2009; Stillman & Johnson, 1994). The calculated gender bias for this book, -0.0097 using Word2Vec and 0.321101 using BERT, does not fully reflect the book's feminist undertones. The limitations may be attributed to gender stereotypes present in word embeddings, leading to biases in the calculated scores. The existing gender biases in word embeddings might not adequately capture the nuances of feminist contexts and themes, potentially resulting in misrepresentations of the book's actual content. Additionally, our method may struggle to comprehend the dual-layered irony inherent in the narrative, where the extreme scenario presented serves as a critique of misogyny rather than an endorsement.

Grammatical Gender and Biases Across Languages

The consideration of grammatical gender in certain languages, and its absence in others, poses challenges when examining and comparing degrees of biases across different linguistic contexts. However, for this paper, our focus is solely on the English language, making this particular concern irrelevant to our study. English, with its use of gendered pronouns, amplifies gender differences, thereby contributing to inherent gender bias and influencing the perceptions of English speakers. While it would be intriguing to investigate how linguistic and grammatical structures in various languages impact gender biases in texts, such an examination is beyond the scope of this research. Future investigations could explore this aspect by comparing word embedding models trained on the same text translated into different languages, thus shedding light on the influence of linguistic factors on gender representation and biases. Nevertheless, such an exploration warrants a separate and dedicated research endeavor.

Detection of Non-Binary Gender

This leads to the third problem of measuring gender. Despite the widely recognized understanding that gender exists along a spectrum (Bamman, 2014b; Richards et al., 2016), existing research papers in the area often categorize gender into binary distinctions of male and female, while neglecting the existence of non-binary genders (Devinney et al., 2022; Stanczak & Augenstein, 2021; Sun et al., 2021). Larson (2017) further argues that the use of gender as a variable in NLP itself is an ethical issue as many research assign gender labels (typically binary) to texts and authors without addressing the methodology behind such assignments and the complexities of gender identity. This method and oversimplification of gender overlook the diverse ways in which individuals experience and express their gender identities, perpetuating gender stereotypes and erasing the experiences of marginalized individuals (Dev et al., 2021; Devinney et al., 2022; Larson, 2017). Thus, researchers must be mindful of the potential impact their research can have on marginalized groups and strive to create inclusive frameworks that accurately reflect the diversity of gender identities.

Addressing these ethical concerns requires adopting a more nuanced approach to gender representation in NLP research. This includes acknowledging the existence of non-binary genders and incorporating inclusive methodologies that allow individuals to self-identify their gender rather than imposing binary labels. Recently, several pieces of research innovated upon existing techniques to include non-binary gender in their system's method. Manzini et al. (2019) generalized the debiasing method established by Bolukbasi et al. (2016) and applied it to multi-class debiasing. Sun et al. (2021) replaced binary pronouns with singular they/them pronouns, often used by non-binary and genderqueer individuals, to train models that generate singular they/them pronouns.

Compared to the two binary genders, the non-binary gender, due to its complex and fluid nature, does not have a singular word or pronoun representative of the entire identity group (Lauscher et al., 2022). The words, such as 'non-binary' and 'genderqueer,' that indicate non-binary gender was used in the 1990s within the Queer community (Matsuno & Budge, 2017) but are, unfortunately, not present in the word embedding we are using in this research, lending the research's use of binary characterization. In the Word2Vec method, our binary characterization indicates how many traditionally female or male-leaning words and descriptions the text contains. BERT method predicts the probability of the text having gender bias. As neither of the methods is concerned with the gender identity of characters in the books and simply uses traditionally binary social gender characteristics as an indicator of potential gender bias within the text, the direct inclusion of the non-binary gender identity requires different approaches and considerations that were beyond the scope of this study.

CONCLUSIONS

Innovation and Contribution

This study employed two NLP methodologies, namely Word2Vec and BERT models, to assess gender bias in popular fiction and analyze consumers' preferences regarding gender bias in titles. Our findings reveal a trend over time, wherein consumers of books and movies gravitate towards content that exhibits a female bias or no gender bias.

Diverging from the predominant focus of prior studies on identifying gender bias in various NLP models and developing techniques to mitigate its adverse impact on downstream tasks (Stanczak & Augenstein, 2021), our research aligns with Garg et al. (2018) in perceiving gender bias in NLP as indicative of societal gender bias and employs it to quantify gender bias in popular fiction.

Additionally, our approach integrates consumer demand for media, or its popularity, when assessing societal gender bias. We argue that highly biased media may have minimal impact if not consumed, whereas slightly biased media, when widely shared, can accurately reflect prevailing societal bias. By accounting for calculated bias in media with consumer demand information, our results provide a more precise representation of societal preferences than solely considering the supply side.

Furthermore, our research expands on prior work by utilizing both word embeddings and BERT models to gauge bias. The use of two methods to evaluate the same dataset allows for comparison and validation of the results produced by each model, thereby reinforcing their effectiveness in measuring gender bias.

In conclusion, this research underscores the value of leveraging NLP techniques, such as Word2Vec and BERT models, in analyzing gender representation in various forms of media. By incorporating both the supply and demand aspects of media consumption, our approach offers a more accurate reflection of societal gender biases, enabling a more informed examination of the evolution of gender representation in popular culture. This study contributes to ongoing efforts aimed at fostering a more inclusive and equitable society, where diverse media representation becomes the norm.

Limitations and Future Works

While our study provides valuable insights, there is room for improvement in three key areas –data, additional parameters, and methodologies. A primary limitation stems from the data used in our analysis. The book data from the Seattle Public Library serves as a case study of Seattle and may not fully represent consumption preferences throughout the entire United States. Being representative of a metropolitan area, the dataset may reflect a liberal-leaning reading preference but lacks sufficient geographical diversity to capture the broader spectrum of the US consumers' book preferences. Moreover, the rise of online media and the declining popularity of libraries as entertainment sources necessitates the inclusion of a wider range of media types, such as YouTube videos or podcasts, to capture diverse demographic preferences accurately. Additionally, the limited accessibility of physical book copies in the library compared to online media may hinder the library's ability to precisely reflect dynamic and peak consumer demand (Baran, 2013; Rajendran & Thesinghraja, 2014).

In the future, our research could benefit from incorporating supplementary variables or parameters, such as the gender of the producer or author, as well as consumer demographics and regional information. This additional data would facilitate more comprehensive comparative analyses of societal bias. For instance, by obtaining checkout records from public libraries across the US, we could generate results that better represent consumer preferences across diverse regions. Furthermore, acquiring further media-related data could shed light on specific experimental outcomes, including the divergent trends of gender bias in songs compared to books and movies.

We also encourage future researchers to explore alternative methods for quantifying textual bias beyond the two approaches utilized in our study. Gender bias and representation in media are multifaceted concepts that affect individuals based on their unique backgrounds and experiences. Although our word embedding and BERT techniques have demonstrated efficacy, they may not fully capture the complexity of the issue. Large language models (LLMs), such as GPT, present promising avenues for detecting gender bias in text. LLMs possess the capability to capture nuanced language patterns and contextual information, facilitating more accurate analyses of gender bias across various media forms. Despite their advantages, LLMs inherit biases from their training data (Nozza et al., 2022; Schramowski et al., 2022), which can lead to biased outputs. Additionally, their lack of interpretability poses challenges to understanding their reasoning (Yang et al., 2023; Zhuo et al., 2023). Moreover, the resource-intensive nature of training LLMs and their technical complexity present challenges for their use in the field of social science. To overcome these obstacles and ensure a comprehensive analysis of gender bias in LLMs, a multidisciplinary approach is needed, combining technical expertise with social sciences knowledge. Advancements should focus on developing more robust and transparent models, complemented by human analysis and interdisciplinary collaboration.

Implications of the Work

The findings of this research hold significant implications for both academic and practical domains, shedding light on the understanding and quantification of gender bias in popular media. By enriching the existing body of knowledge in this area, our study contributes to various fields, including media studies, gender studies, and NLP.

In media studies, our research emphasizes the importance of investigating gender representation and bias across various media types, facilitating a comprehensive understanding of societal preferences' evolution. Within the context of gender studies, our presented methodology allows researchers to track shifts in societal gender bias over time and across diverse media formats, providing insight into the cultural, social, and political factors influencing these trends. In the field of NLP, our work showcases the potential of advanced techniques, such as Word2Vec and BERT models, in quantifying gender bias in texts. This advancement may inspire the development of new algorithms and models tailored to address other forms of bias or representation, thereby expanding the application of NLP to diverse social and cultural issues.

The practical implications of our research extend to media rating systems and content advisory boards. We envision the incorporation of quantified gender bias as a parameter in rating systems, such as the Motion Picture Content Rating System, enabling consumers to make well-informed decisions regarding the gender representation and bias present in media content. By integrating quantified gender bias into these systems, consumers can make more informed choices about the media they consume, aligning with their preferences and values. This development may stimulate a higher demand for unbiased content, potentially fostering a shift towards more balanced and diverse gender representations across various media formats. Additionally, policymakers and advocacy groups can utilize our research to develop content creation guidelines that prioritize inclusive and balanced representations, implement incentives for diversity in media production, and monitor the market for gender bias in content. By employing these strategies, policymakers and advocacy groups can work towards creating a more inclusive and equitable media landscape that challenges stereotypes and promotes diverse perspectives on gender.

This research acknowledges the ethical implications related to gender bias in media representation. By addressing these concerns proactively, we aim to contribute to the ongoing discourse on enhancing gender representation in popular fiction. It is vital to highlight the potential consequences of our research findings, such as the reinforcement of harmful stereotypes or the potential for positive change. Adopting a comprehensive and ethical perspective will ensure a well-rounded understanding of the study's impact and implications.

In summary, our research has implications that span multiple disciplines, impacting media creation, policy-making, and public awareness. By quantifying gender bias in popular fiction and advancing the application of NLP techniques, we strive to contribute to a more inclusive and unbiased representation of gender across various forms of media.

Author contributions: All authors were involved in concept, design, collection of data, interpretation, writing, and critically revising the article. All authors approved the final version of the article.

Funding: The authors received no financial support for the research and/or authorship of this article.

Ethics declaration: The authors declared that the study used secondary data that was available in the public domain. No humans were involved in the study. The study was conducted according to the institutional ethical regulations.

Declaration of interest: Authors declare no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

Abbott, T. B. (2013). The trans/romance dilemma in *Transamerica* and other films. *The Journal of American Culture, 36*(1), 32-41. https://doi.org/10.1111/jacc.12011

Adichie, C. N. (2009). The danger of a single story. *TED*. https://www.ted.com/talks/chimamanda_adichie_the_danger_of_a_single_story

Amossy, R., & Heidingsfeld, T. (1984). Stereotypes and representation in fiction. *Poetics Today, 5*(4), 689-700. https://doi.org/10.2307/1772256

Asr, F. T., Mazraeh, M., Lopes, A., Gautam, V., Gonzales, J., Rao, P., & Taboada, M. (2021). The gender gap tracker: Using natural language processing to measure gender bias in media. *PLoS ONE, 16*(1), 1-28. https://doi.org/10.1371/journal.pone.0245533

Atwood, M. (2017). Margaret Atwood on what 'The handmaid's tale'means in the age of Trump. *The New York Times*. https://www.nytimes.com/2017/03/10/books/review/margaret-atwood-handmaids-tale-age-of-trump.html

Babaeianjelodar, M., Lorenz, S., Gordon, J., Matthews, J., & Freitag, E. (2020). Quantifying gender bias in different corpora. In *Proceedings of the Companion Web Conference 2020*. https://doi.org/10.1145/3366424.3383559

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics, 18*(2), 135-160. https://doi.org/10.1111/josl.12080

Bamman, D., Underwood, T., & Smith, N. A. (2014). A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 370-379). https://doi.org/10.3115/v1/P14-1035

Baran, R. A. (2013). *Re-interpretation of library program: The Seattle Public Library* [Master's thesis, Middle East Technical University].

Basta, C., Costa-Jussà, M. R., & Casas, N. (2019). Evaluating the underlying gender bias in contextualized word embeddings. *ArXiv,1904.08783*. https://doi.org/10.18653/v1/W19-3805

Beauchamp, G. (2009). The politics of The handmaid's tale. *The Midwest Quarterly, 51*(1).

Beltrán, M. (2018). Representation. In M. Kackman, & M. C. Kearney (Eds.), *The craft of criticism: Critical media studies in practice* (pp. 94-106). Routledge. https://doi.org/10.4324/9781315879970-9

Betti, L., Abrate, C., & Kaltenbrunner, A. (2023). Large scale analysis of gender bias and sexism in song lyrics. *EPJ Data Science, 12*, 10. https://doi.org/10.1140/epjds/s13688-023-00384-8

Beukeboom, C., & Burgers, C. (2019). How stereotypes are shared through language: A review and introduction of the social categories and stereotypes communication (SCSC) framework. *Review of Communication Research, 7*. https://doi.org/10.12840/issn.2255-4165.017

Bleich, E., Bloemraad, I., & De Graauw, E. (2015). Migrants, minorities and the media: Information, representations and participation in the public sphere. *Journal of Ethnic and Migration Studies, 41*(6), 857-873. https://doi.org/10.1080/1369183X.2014.1002197

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *ArXiv, 2005.14050*. https://doi.org/10.18653/v1/2020.acl-main.485

Boghrati, R., & Berger, J. (2023). Quantifying cultural change: Gender bias in music. *Journal of Experimental Psychology: General, 152*(9), 2591-2602. https://doi.org/10.1037/xge0001412

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *ArXiv:1607.06520*.

Bonomi, A. E., Altenburger, L. E., & Walton, N. L. (2013). "Double crap!" abuse and harmed identity in fifty shades of grey. *Journal of Women's Health, 22*(9), 733-744. https://doi.org/10.1089/jwh.2013.4344

Booker, M. K., & Clapper, T. H. (1995). Review of the dystopian impulse in modern literature: Fiction as social criticism. *Utopian Studies, 6*(2), 147-149.

Brooks, D. E., & Hébert, L. P. (2006). Gender, race, and media representation. *Handbook of Gender and Communication, 16*, 297-317. https://doi.org/10.4135/9781412976053.n16

Burns, K., Hendricks, L. A., Saenko, K., Darrell, T., & Rohrbach, A. (2019). Women also snowboard: Overcoming bias in captioning models. *arXiv.org*. https://arxiv.org/abs/1803.09797

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183-186. https://doi.org/10.1126/science.aal4230

Castañeda, M. (2018). The power of (mis)representation: Why racial and ethnic stereotypes in the media matter. *Challenging Inequalities: Readings in Race, Ethnicity, and Immigration*, *60*.

Chapman, B. V., Rooney, M. K., Ludmir, E. B., De La Cruz, D., Salcedo, A., Pinnix, C. C., Das, P., Jagsi, R., Thomas Jr, C. R., & Holliday, E. B. (2020). Linguistic biases in letters of recommendation for radiation oncology residency applicants from 2015 to 2019. *Journal of Cancer Education, 37*(4), 965-972. https://doi.org/10.1007/s13187-020-01907-x

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science, 32*(2), 218-240. https://doi.org/10.1177/0956797620963619

Chen, Y., Mahoney, C., Grasso, I., Wali, E., Matthews, A., Middleton, T., Njie, M., & Matthews, J. (2021). Gender bias and under-representation in natural language processing across human languages. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3461702.3462530

Clement, M., Fabel, S., & Schmidt-Stolting, C. (2006). Diffusion of hedonic goods: A literature review. *International Journal on Media Management, 8*(4), 155-163. https://doi.org/10.1207/s14241250ijmm0804_1

Costa-Jussà, M. R. (2019). An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence, 1*, 495-496. https://doi.org/10.1038/s42256-019-0105-5

Crabb, P. B., & Bielawski, D. (1994). The social representation of material culture and gender in children's books. *Sex Roles, 30*, 69-79. https://doi.org/10.1007/BF01420740

Dahlgren, P. (2000). Television and the public sphere: Citizenship, democracy and the media. In *SAGE knowledge*. SAGE. https://doi.org/10.4135/9781446250617

Dai, A. M., Olah, C., & Le, Q. V. (2015). Document embedding with paragraph vectors. *ArXiv:1507.07998*. https://arxiv.org/abs/1507.07998

Delgado-Rodriguez, M., & Llorca, J. (2004). Bias. *Journal of Epidemiology & Community Health, 58*(8), 635-641. https://doi.org/10.1136/jech.2003.008466

Delobelle, P., Tokpo, E. K., Calders, T., & Berendt, B. (2021). Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models. *arXiv, 2112.07447*. https://doi.org/10.18653/v1/2022.naacl-main.122

Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J. M., & Chang, K.-W. (2021). Harms of gender exclusivity and challenges in non-binary representation in language technologies. *ArXiv, 2108.12084*. https://doi.org/10.18653/v1/2021.emnlp-main.150

Devinney, H., Björklund, J., & Björklund, H. (2022). Theories of "gender" in NLP bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2083-2102). https://doi.org/10.1145/3531146.3534627

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv, 1810.04805*.

Dhingra, B., Liu, H., Salakhutdinov, R., & Cohen, W. W. (2017). A comparative study of word embeddings for reading comprehension. *arXiv, 1703.00993*.

Dixon-Fyle, S., Dolan, K., Hunt, V., & Prince, S. (2020). Diversity wins: How inclusion matters. *www.mckinsey.com*. https://www.mckinsey.com/featured-insights/diversity-and-inclusion/diversity-wins-how-inclusion-matters

Dolci, T. (2022). Fine-tuning language models to mitigate gender bias in sentence encoders. In *Proceedings of the IEEE 8th International Conference on Big Data Computing Service and Applications* (pp. 175-176). https://doi.org/10.1109/BigDataService55688.2022.00036

Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., & Tang, J. (2021). All NLP tasks are generation tasks: A general pretraining framework. *ArXiv, 2103.10360*. https://arxiv.org/abs/2103.10360

Fast, E., Vachovsky, T., & Bernstein, M. S. (2016). Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. *ArXiv, 1603.08832*. https://arxiv.org/abs/1603.08832

Fiske, S. T. (1993). Controlling other people: The impact of power on stereotyping. *American Psychologist, 48*(6), 621-628. https://doi.org/10.1037/0003-066x.48.6.621

Friedman, S., Schmer-Galunder, S., Chen, A., & Rye, J. (2019). *Relating word embedding gender biases to gender gaps: A cross-cultural analysis*. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3803

Fryberg, S. A., Markus, H. R., Oyserman, D., & Stone, J. M. (2008). Of warrior chiefs and Indian princesses: The psychological consequences of American Indian mascots. *Basic and Applied Social Psychology, 30*(3), 208-218. https://doi.org/10.1080/01973530802375003

Fürsich, E. (2010). Media and the representation of others. *International Social Science Journal, 61*(199), 113-130. https://doi.org/10.1111/j.1468-2451.2010.01751.x

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences, 115*(16), E3635-E3644. https://doi.org/10.1073/pnas.1720347115

Glick, P., & Fiske, S. T. (2011). Ambivalent sexism revisited. *Psychology of Women Quarterly, 35*(3), 530-535. https://doi.org/10.1177/0361684311414832

Goldfarb-Tarrant, S., Marchant, R., Sánchez, R. M., Pandya, M., & Lopez, A. (2020). Intrinsic bias metrics do not correlate with application bias. *arXiv, 2012.15859*.

Green, M. C., Garst, J., & Brock, T. C. (2003). The power of fiction: Determinants and boundaries. In L. J. Shrum (Ed.), *The psychology of entertainment media*. Erlbaum Psych Press.

Guthrie, D., Allison, B., Liu, W., Guthrie, L., & Wilks, Y. (2006). *A closer look at skip-gram modelling*. European Language Resources Association.

Hagiwara, N., Slatcher, R. B., Eggly, S., & Penner, L. A. (2017). Physician racial bias and word use during racially discordant medical interactions. *Health Communication, 32*(4), 401-408. https://doi.org/10.1080/10410236.2016.1138389

Hall, S. (1997). Culture and power. *Radical Philosophy, 86*(27), 24-41. https://doi.org/10.1177/004839319702700102

Hamilton, M. C., Anderson, D., Broaddus, M., & Young, K. (2006). Gender stereotyping and under-representation of female characters in 200 popular children's picture books: A twenty-first century update. *Sex Roles, 55*(11-12), 757-765. https://doi.org/10.1007/s11199-006-9128-6

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). Diachronic word embeddings reveal statistical laws of semantic change. *ArXiv, 1605.09096*. https://arxiv.org/abs/1605.09096

Hanne, M. (1994). *The power of the story: Fiction and political change*. Berghahn Books.

Harrington, C. (2021). What is 'toxic masculinity' and why does it matter? *Men and Masculinities, 24*(2), 345-352. https://doi.org/10.1177/1097184X20943254

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass, 15*(8), e12432. https://doi.org/10.1111/lnc3.12432

Huang, G., Li, K., & Li, H. (2019). Show, not tell: The contingency role of infographics versus text in the differential effects of message strategies on optimistic bias. *Science Communication, 41*(6), 732-760. https://doi.org/10.1177/1075547019888659

Hubler, A. E. (2000). Beyond the image: Adolescent girls, reading, and social reality. *NWSA Journal, 12*(1), 84-99. https://doi.org/10.2979/NWS.2000.12.1.84

James, S. E., Herman, J., Keisling, M., Mottet, L., & Anafi, M. (2019). *2015 U.S. transgender survey (USTS)*. https://www.icpsr.umich.edu/web/RCMD/studies/37229

Johnson, D. R., Huffman, B. L., & Jasper, D. M. (2014). Changing race boundary perception by reading narrative fiction. *Basic and Applied Social Psychology, 36*(1), 83-90. https://doi.org/10.1080/01973533.2013.856791

Johnson, R. (2008). Assessment of bias with emphasis on method comparison. *The Clinical Biochemist. Reviews, 29*(Suppl 1), S37-S42.

Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the linguistic inquiry and word count. *The American Journal of Psychology, 120*(2), 263-286. https://doi.org/10.2307/20445398

Kearl, H. (2014). Unsafe and harassed in public spaces: A national street harassment report. *ncvc.dspacedirect.org*. https://ncvc.dspacedirect.org/handle/20.500.11990/479

Khadilkar, K., KhudaBukhsh, A. R., & Mitchell, T. M. (2022). Gender bias, social bias, and representation in Bollywood and Hollywood. *Patterns, 3*(4), 100486. https://doi.org/10.1016/j.patter.2022.100486

Khan, U., Dhar, R., & Wertenbroch, K. (2005). A behavioral decision theory perspective on hedonic and utilitarian choice. In D. Mick, & S. Ratneshwar (Eds.), *Inside consumption: Consumer motives, goals, and desires*. Routledge.

Kidd, M. A. (2016). Archetypes, stereotypes and media representation in a multi-cultural society. *Procedia-Social and Behavioral Sciences, 236*, 25-28. https://doi.org/10.1016/j.sbspro.2016.12.007

Kraicer, E., & Piper, A. (2019). Social characters: The hierarchy of gender in contemporary English-language fiction. *Journal of Cultural Analytics, 3*(2). https://doi.org/10.22148/16.032

Kupers, T. A. (2005). Toxic masculinity as a barrier to mental health treatment in prison. *Journal of Clinical Psychology, 61*(6), 713-724. https://doi.org/10.1002/jclp.20105

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From word embeddings to document distances. *proceedings.mlr.press*. https://proceedings.mlr.press/v37/kusnerb15.html

Lai, S., Liu, K., He, S., & Zhao, J. (2016). How to generate a good word embedding. *IEEE Intelligent Systems, 31*(6), 5-14. https://doi.org/10.1109/MIS.2016.45

Lai, Y. A., Lalwani, G., & Zhang, Y. (2020). Context analysis for pre-trained masked language models. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 3789-3804). https://doi.org/10.18653/v1/2020.findings-emnlp.338

Larson, B. (2017). *Gender as a variable in natural-language processing: Ethical considerations*. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1601

Lau, J. H., & Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. *ArXiv, 1607.05368*. https://doi.org/10.18653/v1/W16-1609

Lauscher, A., Crowley, A., & Hovy, D. (2022). Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. *ArXiv, 2202.11923*. https://arxiv.org/abs/2202.11923

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning* (pp. 1188-1196). PMLR.

Leavitt, P. A., Covarrubias, R., Perez, Y. A., & Fryberg, S. A. (2015). "Frozen in time": The impact of native American media representations on identity and self-understanding. *Journal of Social Issues, 71*(1), 39-53. https://doi.org/10.1111/josi.12095

Li, S., Fant, A. L., McCarthy, D. M., Miller, D., Craig, J., & Kontrick, A. (2017). Gender differences in language of standardized letter of evaluation narratives for emergency medicine residency applicants. *AEM Education and Training, 1*(4), 334-339. https://doi.org/10.1002/aet2.10057

Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2020). Gender bias in neural natural language processing. *Logic, Language, and Security, 12300*, 189-202. https://doi.org/10.1007/978-3-030-62077-6_14

Madaan, N., Mehta, S., Agrawaal, T. S., Malhotra, V., Aggarwal, A., Gupta, Y., & Saxena, M. (2018). Analyze, detect and remove gender stereotyping from Bollywood movies. In S. A. Friedler, & C. Wilson (Eds.), *Proceedings of the Conference on Fairness, Accountability and Transparency* (pp. 92-105). PMLR.

Madanikia, Y., & Bartholomew, K. (2014). Themes of lust and love in popular music lyrics from 1971 to 2011. *SAGE Open, 4*(3), 2158244014547179. https://doi.org/10.1177/2158244014547179

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. *ArXiv, 1904.04047*. https://doi.org/10.18653/v1/N19-1062

Mar, R. A. , & Oatley, K. (2008). The function of fiction is the abstraction and simulation of social experience . *Perspectives on Psychological Science, 3* , 173-192 . https://doi.org/10.1111/j.1745–6924. 2008.00073.x

Matsuno, E., & Budge, S. L. (2017). Non-binary/genderqueer identities: A critical review of the literature. *Current Sexual Health Reports, 9*(3), 116-120. https://doi.org/10.1007/s11930-017-0111-8

Matthews, A., Grasso, I., Mahoney, C., Chen, Y., Wali, E., Middleton, T., Njie, M., & Matthews, J. (2021). *Gender bias in natural language processing across human languages*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.trustnlp-1.6

Maudslay, R. H., Gonen, H., Cotterell, R., & Teufel, S. (2020). It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. *arXiv.org*. https://arxiv.org/abs/1909.00871

McInroy, L. B., & Craig, S. L. (2016). Perspectives of LGBTQ emerging adults on the depiction and impact of LGBTQ media representation. *Journal of Youth Studies, 20*(1), 32-46. https://doi.org/10.1080/13676261.2016.1184243

McLaren, J. T., Bryant, S., & Brown, B. (2021). "See me! Recognize me!" An analysis of transgender media representation. *Communication Quarterly, 69*(2), 172-191. https://doi.org/10.1080/01463373.2021.1901759

Merchant, A., Rahimtoroghi, E., Pavlick, E., & Tenney, I. (2020). What happens to BERT embeddings during fine-tuning? *arXiv, 2004.14448*. https://doi.org/10.18653/v1/2020.blackboxnlp-1.4

Merrick, H. (2012). Challenging implicit gender bias in science: Positive representations of female scientists in fiction. *Journal of Community Positive Practices, 12*(4), 744-768.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *ArXiv.org*. https://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *ArXiv.org*. https://arxiv.org/abs/1310.4546

Muhammed, M. (2020). Sexism in Wilde's the picture of Dorian Gray: Linguistic analysis. *Journal of Tikrit University for Humanities, 27*(3), 11-26. https://doi.org/10.25130/jtuh.27.3.2020.24

Nadeem, M., Bethke, A., & Reddy, S. (2020). StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv, 2004.09456*.

Nozza, D., Bianchi, F., & Hovy, D. (2022). Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode# 5--Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.bigscience-1.6

Ochieng, D. (2012). Sexism in language: Do fiction writers assign agentive and patient roles equally to male and female characters? *Journal of Language and Linguistic Studies, 8*(2), 0-47.

Olson, D. (2012). From utterance to text: The bias of language in speech and writing. *Harvard Educational Review, 47*(3), 257-281. https://doi.org/10.17763/haer.47.3.8840364413869005

Otterbacher, J. (2015). Crowdsourcing stereotypes: Linguistic bias in metadata generated via GWAP. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1955-1964). https://doi.org/10.1145/2702123.2702151

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: Liwc 2001*. Lawrence Erlbaum Associates.

Pennington, J., Socher, R., & Manning, C. (2014). *GloVe: Global vectors for word representation*. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Phillips, J. (2006). Introduction. In *Transgender on screen*. Palgrave Macmillan. https://doi.org/10.1057/9780230596337_1

Rajendran, L., & Thesinghraja, P. (2014). The impact of new media on traditional media. *Middle-East Journal of Scientific Research, 22*(4), 609-616.

Rey, V. (2020). *The art of minorities: Cultural representation in museums of the Middle East and North Africa*. Edinburgh University Press. https://doi.org/10.3366/edinburgh/9781474443760.001.0001

Rezaeinia, S. M., Rahmani, R., Ghodsi, A., & Veisi, H. (2019). Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems With Applications, 117*, 139-147. https://doi.org/10.1016/j.eswa.2018.08.044

Richards, C., Bouman, W. P., Seal, L., Barker, M. J., Nieder, T. O., & T'Sjoen, G. (2016). Non-binary or genderqueer genders. *International Review of Psychiatry, 28*(1), 95-102. https://doi.org/10.3109/09540261.2015.1106446

Said, E. W. (2016). Orientalism. In *Social theory re-wired*. Routledge.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513-523. https://doi.org/10.1016/0306-4573(88)90021-0

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence, 4*(3), 258-268. https://doi.org/10.1038/s42256-022-00458-8

Sculos, B. W. (2017). Who's afraid of 'toxic masculinity'? *Class Race Corporate Power, 5*(3), 6. https://doi.org/10.25148/CRCP.5.3.006517

Shrum, L. J. (2009). Media consumption and perceptions of social reality: Effects and underlying processes. In J. Bryant and M. B. Oliver (Eds.), *Media effects: Advances in theory and research* (pp. 50-73). Routledge.

Smelik, A. (2007). Feminist film theory. In *The cinema book* (pp. 491-504). https://doi.org/10.5040/9781838710484.0065

Smiler, A. P., Shewmaker, J. W., & Hearon, B. (2017). From "I want to hold your hand" to "promiscuous": Sexual stereotypes in popular music lyrics, 1960-2008. *Sexuality & Culture, 21(4)*, 1083-1105. https://doi.org/10.1007/s12119-017-9437-7

Smith, S. L., & Granados, A. D. (2009). Content patterns and effects surrounding sex-role stereotyping on television and films. In J. Bryant and M. B. Oliver (Eds.), *Media effects: Advances in theory and research* (pp. 342-361). Routledge.

Stanczak, K., & Augenstein, I. (2021). A survey on gender bias in natural language processing. *ArXiv, 2112.14168*. https://doi.org/10.48550/arXiv.2112.14168

Stillman, P. G., & Johnson, S. A. (1994). Identity, complicity, and resistance in The handmaid's tale. *Utopian Studies, 5*(2), 70-86.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? In *Proceedings of the Chinese Computational Linguistics: 18th China National Conference* (pp. 194-206). Springer. https://doi.org/10.1007/978-3-030-32381-3_16

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). *Mitigating gender bias in natural language processing: Literature review*. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1159

Sun, T., Liu, X., Qiu, X., & Huang, X. (2022). Paradigm shift in natural language processing. *Machine Intelligence Research, 19*(3), 169-183. https://doi.org/10.1007/s11633-022-1331-6

Sun, T., Webster, K., Shah, A., Wang, W. Y., & Johnson, M. (2021). They, them, theirs: Rewriting with gender-neutral English. *ArXiv, 2102.06788*. https://arxiv.org/abs/2102.06788

Sutton, A., Lansdall-Welfare, T., & Cristianini, N. (2018). Biased embeddings from wild data: Measuring, understanding and removing. In *Proceedings of the 17th International Symposium* IDA (pp. 328-339). Springer. https://doi.org/10.1007/978-3-030-01768-2_27

Thomas, D. C., Lawlor, D. A., & Thompson, J. R. (2007). Re: Estimation of bias in nongenetic observational studies using "mendelian triangulation" by Bautista et al. *Annals of Epidemiology, 17*(7), 511-513. https://doi.org/10.1016/j.annepidem.2006.12.005

Underwood, T., Bamman, D., & Lee, S. (2018). The transformation of gender in English-language fiction. *Journal of Cultural Analytics, 3*(2). https://doi.org/10.22148/16.019

Van Reenen, D. (2014). Is this really what women want? An analysis of fifty shades of grey and modern feminist thought. *South African Journal of Philosophy, 33*(2), 223-233. https://doi.org/10.1080/02580136.2014.925730

Vijay, D. (2019). Crazy rich Asians: Exploring discourses of orientalism, neoliberal feminism, privilege and inequality. *Markets, Globalization & Development Review, 4*(3). https://doi.org/10.23860/mgdr-2019-04-03-04

Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proceedings of the International AAAI Conference on Web and Social Media* (pp. 454-463). https://doi.org/10.1609/icwsm.v9i1.14628

Waisbord, S. (2004). Media and the reinvention of the nation. In *The SAGE handbook of media studies*. SAGE. https://doi.org/10.4135/9781412976077.n19

Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C.-C. . J. (2019). Evaluating word embedding models: Methods and experimental results. *APSIPA Transactions on Signal and Information Processing, 8*. https://doi.org/10.1017/ATSIP.2019.12

West, J. B. (2010). *Gender bias and stereotypes in young adult literature: A content analysis of novels for middle school students* [Master's thesis, University of North Carolina at Chapel Hill].

Wu, L., Yen, I. E. H., Xu, K., Xu, F., Balakrishnan, A., Chen, P.-Y., Ravikumar, P., & Witbrock, M. J. (2018). Word mover's embedding: From Word2Vec to document embedding. *ArXiv, 1811.01713*. https://doi.org/10.18653/v1/D18-1482

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., & Hu, X. (2023). Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. *arXiv, 2304.13712*.

Zastrow, C., Kirst-Ashman, K. K., & Hessenauer, S. L. (2019). *Empowerment series: Understanding human behavior and the social environment*. Cengage Learning.

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Systems With Applications, 38*(3), 2758-2765. https://doi.org/10.1016/j.eswa.2010.08.066

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2979-2989). https://doi.org/10.18653/v1/d17-1323

Zhao, Y. (2019). Crazy rich Asians: When representation becomes controversial. *Markets, Globalization & Development Review, 4*(3). https://doi.org/10.23860/mgdr-2019-04-03-03

Zhou, Y., & Srikumar, V. (2021). A closer look at how fine-tuning changes BERT. *arXiv, 2106.14282*. https://doi.org/10.18653/v1/2022.acl-long.75

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Exploring AI ethics of ChatGPT: A diagnostic analysis. *arXiv, 2301.12867*.